Introductory Notes on Probability

Nick Roy

shamelessly ~~stolen~~ borrowed with permission from the lovely & brilliant Lea Duncker

Fall 2018

This set of notes is intended to give a quick introduction to basic probability theory and statistical inference. The goal is to convey key concepts and intuition, without an overly technical treatment.

1 Basic Concepts and Definitions

In this section we will briefly introduce some key concepts you will likely come across. Aside form these notes, the introductory material in Bishop's *Pattern Recognition and Machine Learning* (otherwise known as the PNI Bible), MacKay's *Information Theory, Inference, and Learning Algorithms*, or Murphy's *Machine Learning: A probabilistic perspective* are good starting points. These books are also useful resources more generally.

1.1 Probabilities and Random Variables

Probabilites and random variables are the most fundamental concept in probability theory. Intuitively, a random variable is a variable that takes on some values, but we are unsure about what those values might be. Thus, we assign a probability to each of the possible values the random variable could take. We might express this as

$$X \sim p(X)$$

which means the random variable X is distributed according to $p(X)$, where $p(X)$ is a probability distribution. Depending on whether X takes values in the discrete or continuous domain, $p(X)$ is called a probability mass function (pmf) or probability density function (pdf), respectively.

For example, we could consider the random variable $H \in \{0, 1\}$ denoting whether the outcome of a coin toss is heads.

If the coin is fair, we would expect that

$$\lim_{N\to\infty} \frac{\#\text{Heads}}{N} = 0.5$$

as we repeat many coin tosses, where N is the total number of coin tosses. This would be a definition of probability in terms of frequencies: a probability is the relative occurrence of an event over a number of repeated experiments in the limit of infinite repetitions.

Thus, this gives us a pmf for the random variable H:

$H = h$	0	1
$p(H = h)$	0.5	0.5

This frequency-based definition makes sense when thinking about coin tosses, but what about more abstract events? We can't repeat things like a natural disaster infinitely many times. However, we might still want to attach a probability to the occurrence of such an event. Thus, another way to think of probabilities is as beliefs. You could view these beliefs as a state of knowledge, personal or objective, which you can use to reason logically and coherently about the world. This interpretation is known as the Bayesian view of probability, while the former is known as the Frequentist view.

In order for our beliefs to be consistent, we require them to satisfy certain properties. These are also known as the three axioms of probability:

1. Probabilities are non-negative and real: $P(X) \geq 0, \in \mathbb{R}^+$
2. Probabilities are normalized: if discrete $\sum_i P(X_i) = 1$, if continuous $\int P(X)dX = 1$
3. Probabilities of disjoint (mutually exclusive) sets (e.g. alternatives) add: $P(A \cup B) = P(A) + P(B)$, if $P(A \cap B) = 0$.

1.2 Marginals, Joints and Conditionals

So far we have looked at a distribution over a single variable. This is also called the marginal distribution. However, we might also be interested in asking questions about multiple random variables. For example, I might have a belief about how tall a person is. How is that related to that person's weight? Knowing that someone is really tall also changes my belief about how much they weigh. Similarly, knowing someone's weight also gives me some idea about how tall they might be. We can formalise this intuiting using joint and conditional distributions.

A joint distribution is a distribution over two events co-occuring, i.e. the probablity that X and Y occur. We write this as $P(X, Y)$. The conditional distribution of X given Y is defined as

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

and allows us to make statements about our belief about one event given that we know the outcome of another one.

From this definition follows the product rule which rearranges the about to show that the joint distribution can be factorised into

$$P(X, Y) = P(X|Y)P(Y)$$

Another important concept is that of marginalisation. This means that if we sum/integrate over all possible outcomes of one variable in the joint distribution, we obtain the marginal distribution of the other variable:

$$P(X) = \int P(X, Y) dY$$

for continuous Y or, for discrete Y:

$$P(X) = \sum_i P(X, Y_i)$$

1.3 Independence

Independence is an important concept in probability theory. Formally, two random variables are independent if

$$P(X, Y) = P(X)P(Y)$$

From this and the product rule, we also know that this means

$$P(X|Y) = P(X)$$

So knowing Y doesn't tell me anything about X and vice versa – the two variables are independent. Independence means the joint distribution factorises[1].

1.4 Expected Value

The expected value is an important concept. Its formal definition for a random variable $X \sim P(X)$ is

$$\mathbb{E}_{p(X)}[X] = \int x p(x) dx$$

or

$$\mathbb{E}_{p(X)}[X] = \sum_{x_i} x_i p(x_i)$$

for continuous and discrete variables, respectively. This gives us the average value that X would take under the density $P(X)$. For discrete variables this is really intuitive: we just compute a weighted sum of all possible events, where the weight is determined by the probability of that event occurring. We can also take expected values of functions of X:

$$\mathbb{E}_{p(X)}[f(X)] = \int f(x) p(x) dx$$

For $f(X) = (X - \mathbb{E}_{p(X)}[X])^2$ this gives us the variance of X under $p(X)$.

[1] Don't think about independence in terms of correlations: uncorrelated variables are not necessarily independent!

1.5 Bayes Theorem

Bayes Theorem is

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Bayes theorem basically tells us that we can learn something about an underlying, unknown cause by recording data that was generated by that cause, or in other words, $X|Y$ tells us something about $Y|X$.

1.6 Example: HIV tests

Let us apply Bayes theorem to a simple problem. Suppose you are interested in HIV tests. You know a few things:

- 95% of people who have HIV test positive
- 98% of people who do not have HIV test negative
- the probability of HIV is one in thousand[2]

Suppose you know that someone tested positive. What is your belief about that person actually having HIV?[3]

We can use Bayesian reasoning to answer this, so let us start by translating the information above into numbers. Let $X = 1$ denote that someone has HIV and let $Y = 1$ denote a positive test outcome. We have been given the information

$$\begin{aligned} P(Y = 1|X = 1) &= 0.95 \\ P(Y = 0|X = 0) &= 0.98 \\ P(X = 1) &= 0.001 \end{aligned}$$

The quantity we are interested in is

$$P(X = 1|Y = 1) = \frac{P(Y = 1|X = 1)P(X = 1)}{P(Y = 1)}$$

We have everything we need to calculate the numerator, but we need to work a bit harder to get an expression for the denominator. Luckily, we know about marginalisation, so we can write

$$P(Y = 1) = P(Y = 1, X = 1) + P(Y = 1, X = 0)$$

Applying the product rule, we can write

$$P(Y = 1) = P(Y = 1|X = 1)P(X = 1) + P(Y = 1|X = 0)P(X = 0)$$

[2] this is probably not very accurate but it's just an example

[3] This example is adapted from Jinghao Xue's UCL course on Applied Bayesian Methods

To get expressions for the probabilities in the second term, we need to remember that probabilities sum to 1. Thus

$$P(X = 0) = 1 - P(X = 1) = 0.999$$
$$P(Y = 1|X = 0) = 1 - P(Y = 0|X = 0) = 0.02$$

Plugging everything back into the initial expression we get

$$\begin{aligned} P(X = 1|Y = 1) &= \frac{P(Y = 1|X = 1)P(X = 1)}{P(Y = 1|X = 1)P(X = 1) + P(Y = 1|X = 0)P(X = 0)} \\ &= \frac{0.95 \times 0.001}{0.95 \times 0.001 + 0.02 * 0.999} = 0.0454 \end{aligned}$$